




The UAV Benchmark: Compact Detection of Vehicles in Urban Scenarios

Haitao Lv , Xianwei Zheng , *Member, IEEE*, Xiao Xie , Xueye Chen, and Hanjiang Xiong

Abstract—Vehicle detection in unmanned aerial vehicle (UAV) images is a fundamental task in photogrammetry and remote sensing. While great success has been achieved, this task remains challenging due to two aspects: 1) the limitation of existing annotation methods in compactly enclosing targets with large perspective distortions in oblique UAV images; 2) the lack of vehicle detection datasets under oblique perspectives. To this end, we propose an oblique UAV benchmark for the precise expression and localization of distorted vehicles in urban scenarios. The benchmark consists of 1) a new parallelogramlike bounding box (PBB) annotation for compactly representing vehicles in oblique UAV images; and 2) a large-scale UAV dataset (namely PARA) for vehicle detection with PBB representation. Our PBB representation frees the angle flexibility to allow a compact depiction of vehicles under various perspective distortion, thus overcoming the inherent limits of rectangular representation [like horizontal bounding box (HBB)] used in traditional annotation methods. PARA comprises 1025 high-resolution images and 117 122 manually annotated object bounding boxes obtained from different UAV platforms. The annotated images are collected from scenarios with complex urban backgrounds and different shooting angles to reflect real-world conditions. Moreover, we compared detection algorithms based on the mainstream HBB and PBB representations on the PARA dataset and established a baseline for UAV oblique image-based vehicle detection. Experimental results validate the effectiveness of PBB representation and highlight the challenges posed by PARA.

Index Terms—Benchmark testing, object detection, remote sensing, unmanned aerial vehicle (UAV) image.

I. INTRODUCTION

THE use of unmanned aerial vehicles (UAVs) to acquire high-resolution images has become an indispensable supplement to satellite remote sensing. In recent years, growing interest has turned to the detection of objects in UAV imagery,

due to some attractive properties of UAVs, e.g., high flexibility, various views, and the ability to acquire both the top and side information of objects. Detecting vehicles from UAV images facilitates a variety of modern urban applications, ranging from crowd detection [1], [2], surveillance [3], [4], traffic monitoring [5], [6], search, and rescue [7], [8], etc. However, vehicles in UAV images are generally captured in an oblique view, which often suffer from drastic perspective deformation [9]. Hence, the detection of UAV vehicles faces not only challenges of arbitrary orientation that are common in ground or satellite images, but also the issue of notable shape and appearance distortion of objects.

Over the past decades, the rapid development of deep learning and its success in computer vision have attracted increased attention to precisely locating and representing vehicles in aerial images [10], [11], [12], [13], [14], [15], [16]. In the early stages, many deep detectors adopt horizontal bounding box (HBB) for vehicle detection in natural imagery due to its simplicity and low cost. These detectors can generally yield satisfactory results in scenes with sparse objects and a relatively simple background. However, when applying these detectors to remote sensing scenes where the object instances are densely crowded and arbitrary-oriented, especially in urban areas with high occlusion, their performances are often dramatically degraded. Inspired by oriented text detection benchmarks [17], [18], the oriented bounding box (OBB) was then introduced to address the challenge of detecting crowded objects in aerial images. OBB employs an additional parameter Θ to the HBB representation to describe the orientation of remotely sensed objects in aerial images.

Nevertheless, unlike the nearly vertical shooting angles of orthographic aerial images, the oblique views of UAV images inevitably bring large geometric distortions to the captured objects. As a result, rectangle-based annotation methods like HBB and OBB are incapable of precisely and compactly enclosing the vehicles in UAV images and often introduce extraneous background noise into learned regional features. A comparison between different annotation methods is shown in Fig. 1(a) and (b). It is clear that the rectangle-based annotation methods fail to correctly delineate the shape of vehicles and include much redundant information into the bounding boxes in oblique UAV image. Although some researchers [19] have employed mask segmentation to achieve more precise vehicle representations in low-altitude UAV images, this annotation significantly increases costs in terms of target annotations and network parameters.

Manuscript received 7 May 2024; revised 4 July 2024; accepted 27 July 2024. Date of publication 14 August 2024; date of current version 5 September 2024. This work was supported in part by the National Natural Science Foundation of China Project under Grant 42071370 and in part by the Fundamental Research Funds for the Central Universities of China under Grant 2042022dx0001. (Corresponding authors: Xianwei Zheng; Xiao Xie.)

Haitao Lv, Xianwei Zheng, and Hanjiang Xiong are with the State Key Laboratory LIESMARS, Wuhan University, Wuhan 430079, China (e-mail: lht0310@whu.edu.cn; zhengxw@whu.edu.cn; xionghanjiang@whu.edu.cn).

Xiao Xie is with the Key Lab for Environmental Computation and Sustainability of Liaoning Province, Institute of Applied Ecology, Chinese Academy of Sciences, Shenyang 110016, China (e-mail: xiexiao@iae.ac.cn).

Xueye Chen is with the The Key Laboratory of Urban Land Resources Monitoring and Simulation, Ministry of Natural Resources, Shenzhen 518000, China (e-mail: xueye31@163.com).

Data is available on-line at <https://github.com/Geo-Tell/PARA>.

Digital Object Identifier 10.1109/JSTARS.2024.3443268

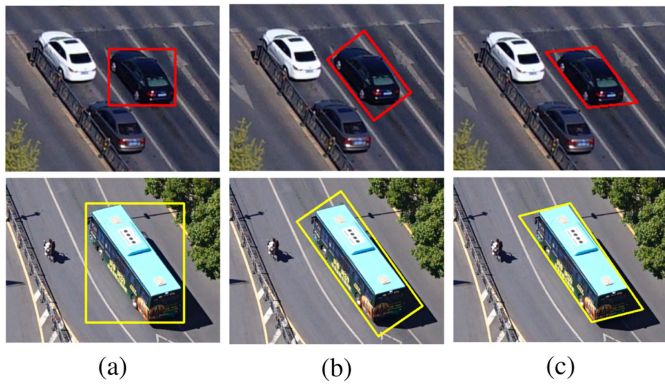


Fig. 1. Comparison between different vehicle annotation methods. (a), (b), and (c) represent HBB, RBB, and our PBB annotation of the same vehicle in the oblique UAV image, respectively.

Except for the inaccurate representation of vehicles, the lack of available datasets for training vehicle detectors for oblique UAV images is another obstacle in this research field. Although several vehicle detection datasets have been developed on aerial images [20], [21], [22], [23], [24], [25], [26], they still have difficulty meeting the requirements of practical applications. In the early period, vehicle datasets [20], [21] had a limited number of instances and low spatial resolution, which restricted their applicability to the detection algorithms. With the fast development of sensor technology, vehicle datasets such as DLR-3K [22], HighD [23], and CARPK [26] began to focus on single or simple natural scenes with high-resolution images. The contained images were collected by low-altitude UAVs on highways or parking lots. However, most of their annotated images are captured in ideal conditions (clear simple backgrounds and without crowded instances), which are inadequate to reflect the complex real-world scenes. To remedy this problem, several large-scale vehicle datasets, such as COWC [24], UCAS-AOD [25], and HRRSD [27], were proposed, which involve more complex backgrounds and a larger number of targets. Nevertheless, these datasets often observe objects from a nearly vertical view, making it difficult to reflect the characteristics of objects in the oblique view images. In contrast to the single perspective of natural and aerial images, the views of UAVs could be varied when the shooting pose changes. As a result, the objects in oblique UAV images frequently suffer from large perspective distortion, posing a significant challenge to the detection algorithms. Due to the lack of annotated UAV images, many methods [28], [29], [30] rely on transfer learning with large-scale natural image datasets (such as ImageNet [31], COCO [32], and Microsoft VOC [33]) for vehicle detection in UAV imagery. Unfortunately, the bias between the UAV datasets and natural datasets makes it hard to learn useful features for UAV objects from natural samples.

Based on the above analysis, we propose a new parallelogramlike representation, namely, parallelogram bounding box (PBB), to compactly enclose vehicles in UAV images. As shown in Fig. 1(c), our PBB can fit well with the shape and orientation of vehicles in UAV images. As a result, the image region contained by our PBB can better reflect the appearance and size of vehicles

under the oblique views compared with conventional annotation methods. This will bring benefits to the learning of more target-related features for vehicle detection in UAV images, as the interference of backgrounds is greatly reduced. We also observe that rectangular vehicles struggle to keep axis-aligned under the oblique views and the parallelogramlike representation can effectively encode the centrosymmetric shape of objects in UAV imagery.

To facilitate the precise detection of vehicles from oblique UAV images with PBB, we further propose a corresponding UAV dataset called PARA for training deep detectors. We collected 1025 UAV images from complex urban scenarios captured by different sensors and platforms. The images in PARA contain vehicles of different appearances, scales, and orientations, with sizes ranging from 1280×1280 pixels to 4000×6000 pixels. All images in PARA are manually annotated by experts in image interpretation with a total of 117 122 PBBs and HBBs. In addition, PARA has some distinctive properties: 1) containing massive vehicle instances under oblique perspectives, which effectively supplements the lack of samples under the multiview observations; and 2) using a novel annotation method that yields a compact vehicle representation that accurately reflects the size and shape of vehicles. We evaluate several mainstream detection algorithms on PARA to illustrate the effectiveness of PBB and build a baseline for vehicle detection in oblique UAV images. The contributions of this work mainly lie in three aspects.

- 1) We proposed a new PBB representation for compact and precise vehicle detection from oblique UAV images. PBB can better fit the shape and orientation of vehicle objects under large perspective deformation than existing HBB and RBB representation.
- 2) We built an UAV dataset (PARA) to facilitate the detection of vehicles with PBB representation, which contains a large amount of images reflecting the oblique nature of UAV data in real-world conditions. In addition, we also divided vehicles into two categories—dynamic vehicles and static vehicles as a complement to the existing vehicle datasets.
- 3) We evaluated several mainstream object detection methods on PARA and proposed a baseline detector, which can provide references for subsequent research on vehicle detection.

II. RELATED WORK

A. Vehicle Detection Method

Vehicle detection is a fundamental subtask in object detection, which aims at locating and classifying vehicles in various types of scene images. With the rapid development of modern detectors, vehicle detection has made significant progress in recent years. To date, the most widely used detectors are deep learning-based ones, which can be divided into two-stage detectors and one-stage detectors. The two-stage detectors first generate candidate regions and then extract the regional features to regress the final bounding box for the target object. These detectors are typically built upon CNN-based models, like RCNN [34], Fast

RCNN [35], and Faster RCNN [36]. To enhance the ability of the network for detecting small targets, FPN [37] aggregates multiscale semantic information using a feature pyramid network to enhance the robustness of the detectors. Cascade RCNN [38] utilizes a multistage network with different IoU thresholds to achieve more precise detection.

Unlike the two-stage detectors, the one-stage detectors treat object detection as a regression problem and use a single-stage network to directly predict the class and position of the target object. The representative one-stage detectors are YOLO series [39], [40], [41], [42], and SSD [43]. Other detectors, like DSSD [44] and FSSD [45], explore how to better fuse multiscale features to improve the detection accuracy of small targets. To solve the problem of extreme imbalance between foreground and background bounding boxes in object detection, RetinaNet [46] introduces focal loss to force the network to pay more attention to hard samples.

Recent advancements in remote sensing have made aerial images with wide coverage and a large number of ground objects widely available. Many studies are devoted to detecting objects with arbitrary orientations in aerial images [47], [48], [49], [50], [51], [52], [53], [54], [55]. In the early stages, many conventional detection methods were developed and modified to predict rotated objects in aerial images. For instance, Rotated Faster RCNN [47] and R2CNN [48] adjust the original Faster RCNN [36] to predict the rotated bounding box (RBB) of the aerial objects. However, limited by the original RPN network, which can only generate horizontal candidate regions, most detectors cannot achieve satisfactory performance in rotated object detection. To solve this problem, several works have been presented to modify the original RPN network [49], [50]. In this line of research, RRPN [18] generates prior rotated boxes of various sizes and aspect ratios onto the feature maps and then feeds prior boxes into the rotated RPN network to generate high-quality rotated candidate regions. ROI transformer [51] devises an ROI transformer module to transform horizontal ROIs into rotated ROIs, thus avoiding producing a large number of anchors and alleviating misalignment problems. Oriented RCNN [53] uses an oriented RPN network to directly generate rotation ROIs, which eliminates the accuracy loss incurred by transforming horizontal ROIs into oriented ROIs.

B. Aerial UAV Datasets

Over the past decades, several UAV datasets have been proposed and employed in various tasks such as object counting, detection, and tracking. Robicquet et al. [56] proposed the STANFORD CAMPUS, which collects image and video data from eight scenes at the Stanford campus. The dataset includes six common categories, such as pedestrian, car, and so on. Zhu et al. [57] introduced the VISDRONE dataset, a high-resolution UAV dataset that comprises more than 200 frames of video and 10 209 UAV images. This dataset provides rich auxiliary data such as bounding boxes, categories, and occlusion ratios. Bozcan et al. [58] proposed the first outdoor multimodal UAV dataset for object detection, i.e., AU-AIR,

which contains target location and attribute information as well as UAV flight statistic data. Du et al. [59] proposed a large-scale UAV image dataset named UAVDT for target detection and tracking. The dataset includes more than 80 000 key frames selected from a 10-hour video, consisting of roughly 2700 vehicles annotated by approximately 840 000 bounding boxes for detection and tracking. Lyu et al. [60] proposed UAVid, a high-resolution UAV dataset for semantic segmentation. UAVid is composed of 300 UAV images taken from 30 video sequences captured in urban areas, with annotations classified by eight categories.

To promote the development of small object detection, Akshatha et al. [61] proposed a large-scale UAV pedestrian detection dataset, namely, Manipal-UAV. This dataset includes 33 videos captured by UAVs at the flight height range of 10–50 m. They selected 13 462 images and annotated 153 112 pedestrian targets. Matthias et al. [62] proposed UAV123, a low-altitude UAV dataset for target tracking. This dataset aims to identify different types of objects and serve applications such as target tracking and trajectory prediction. Wang et al. [63] proposed UAVBD, a low-altitude UAV dataset aimed at detecting abandoned plastic bottles in the wild. This dataset comprises 25 407 UAV images with different backgrounds and 34 791 rotating bounding boxes for bottles. Du et al. [64] proposed UA-DETRAC, a large-scale UAV image dataset for multiobject tracking. They manually annotated 8250 bounding boxes of vehicles in Beijing and Tianjin and provided auxiliary information such as location, illumination, and shooting angles.

C. Vehicle Detection Datasets

In the last decade, there has been an increased attention focused on real-world reflection within datasets, with vehicles being a common object studied in research. In the early stage, datasets such as TAS [20] and OIRDS [21] facilitated the advancement of automated vehicle detection, by employing vehicles as detection categories in satellite remote sensing images. However, the low image resolution of these datasets brings difficulty to accurately reflecting real-world scenarios. With the development of sensor technology, vehicle detection datasets with high-resolution images such as UCAS-AOD [25], VEDAI (2015), and DLR-3K [22] were proposed. Nevertheless, their limited sample quantities hampered their practical application. Similarly, COWC [24] provides a large number of detection targets, 32 716 in total, but its low image resolution and center point-based annotation method impeded the applications of detection algorithms.

The advent of deep learning has resulted in a higher demand for large-scale datasets. Consequently, CPRPK [26] built a large-scale aerial dataset for vehicle detection and counting. It was collected from a parking lot by a drone, with a total of 89 777 vehicle targets. The limitation of dataset is that the scenes in CARPK are too similar to reflect the complexity of the real world. The HighD [23] dataset used drones to capture orthographic images above German highways, ranging from 100 to several hundred meters in elevation. However, its utility is

TABLE I
COMPARISON BETWEEN PARA AND OTHER VEHICLE DATASETS

Dataset	Annotation method	Vehicle Categories	Vehicles	Photograph	Images	Image Width(px)	Year
TAS [20]	HBB	1	1310	orthographic	30	792	2008
UCAS-AOD [25]	HBB	1	2819	orthographic	310	1280	2015
VEDAI [65]	RBB	6	6655	orthographic	1210	1024	2015
DLR-3K [22]	RBB	2	14 235	orthographic	20	5616	2015
COWC [24]	DOT	1	32 716	orthographic	53	2000–19 000	2016
CPRPK [26]	HBB	1	89 777	orthographic	1448	1280	2017
DOTA [47]	RBB	2	43 262	orthographic	2806	800–13 000	2018
DIOR [66]	HBB	1	40 000	orthographic	23 463	800	2019
MHOR [12]	HBB	2	37 806	orthographic	10 631	5000–8000	2020
EAGLE [67]	RBB	2	215 986	orthographic	8280	936	2021
DroneVehicle [68]	RBB	5	953 087	oblique	56 878	840	2022
PARA (ours)	PBB	2	88 396	oblique	1,025	1000–6000	2023

BB, short for the bounding box. DOT stands for using the center point as The object representation. Photograph denotes the photographing method for most images in the datasets.

restricted by its simple background and its inability to apply to complex scenes.

Recently, many datasets have been dedicated to reflecting complex scenes of the real world, which contain more complex background information and instances. For example, DOTA [47] is a large-scale dataset for object detection in aerial images, mainly containing 2806 aerial images captured from different sensors and platforms. DOTA provides the ability to evaluate object detection and rotated object detection in aerial images. Vehicles are considered as a major detection category in this dataset, with a total of 43 462 objects. MOHR [12] collected 10 631 UAV images from suburban areas, including 12 602 trucks and 25204 cars annotated for detection evaluation. VAID [69] collected 6000 aerial images under different lighting conditions in Taiwan. It classifies vehicles into seven categories, like sedan, minibus, truck, pickup, bus, cement truck, and trailer. Currently, the largest dataset for vehicle detection in aerial images is EAGLE [67], which involves 8820 aircraft aerial images shot under various weather, lighting, and humidity conditions. EAGLE contains a total of 215 986 detection targets, including 208 963 small vehicles and 7023 large vehicles.

Although the above datasets cover many real-world scenarios, few datasets pay attention to the influence of shooting angles on the shape and appearance of vehicles. They often use a single vertical view, ignoring the influences caused by the various oblique views. In contrast, the proposed PARA contains a large number of oblique view images and uses a novel object annotation, PBB, to compactly enclose the vehicles in the oblique UAV imagery.

III. PARA DATASET

In this section, we will introduce the proposed PARA dataset, including the source of images, the selection of categories, and the specially designed annotation method. We also make a comprehensive comparison between PARA and other related benchmark datasets in vehicle detection, which is presented in Table I.

A. Image Collection

PARA dataset aims to reflect the complex urban scenarios with the UAV images taken from various different views, and thus enhance the generalization ability of current detection methods. To this end, we collected 1025 UAV images from a variety of diverse urban scenes, including urban roads, parking lots, crossings, building, and highways. For clarity, some original images in PARA are shown in Fig. 2. All the images are captured by different camera-equipped drones, such as DJI Air 3, under different illumination, resolution, and background to increase the diversity of PARA. Moreover, to avoid a single vertical downward viewing angle, we ensure that the drones collect images with different flight heights and observation angles. The varying flight and views allow our dataset to cover a wide range of real-world scenes vehicles differing in several aspects.

B. Category Selection

In PARA dataset, we divide vehicles into two categories, i.e., static vehicle and dynamic vehicle, to enable the flexibility of our dataset for severing different applications. This is also a supplement to the existing datasets for vehicle detection. Existing vehicle datasets often choose several common categories (e.g., large vehicles, small vehicles, etc.) based on the size of vehicles. Such categories can meet the needs of basic applications such as vehicle counting, object detection and so on. However, it is struggling to serve specific applications. For example, traffic monitoring and management are highly complex tasks due to the drastic increase in the number of vehicles, which need to figure out whether vehicles are moving. The limited categories of existing vehicle datasets make them hard to solve the traffic-related problems, which are common in modern urban applications. Therefore, we divide the annotated vehicles in the PARA dataset into static vehicles and dynamic vehicles. They are labeled according to whether the kind of vehicles is moving and this is judged by experts in UAV image interpretation. Moreover, to ensure the diversity of categories in the PARA dataset, we also include pedestrians as a category in

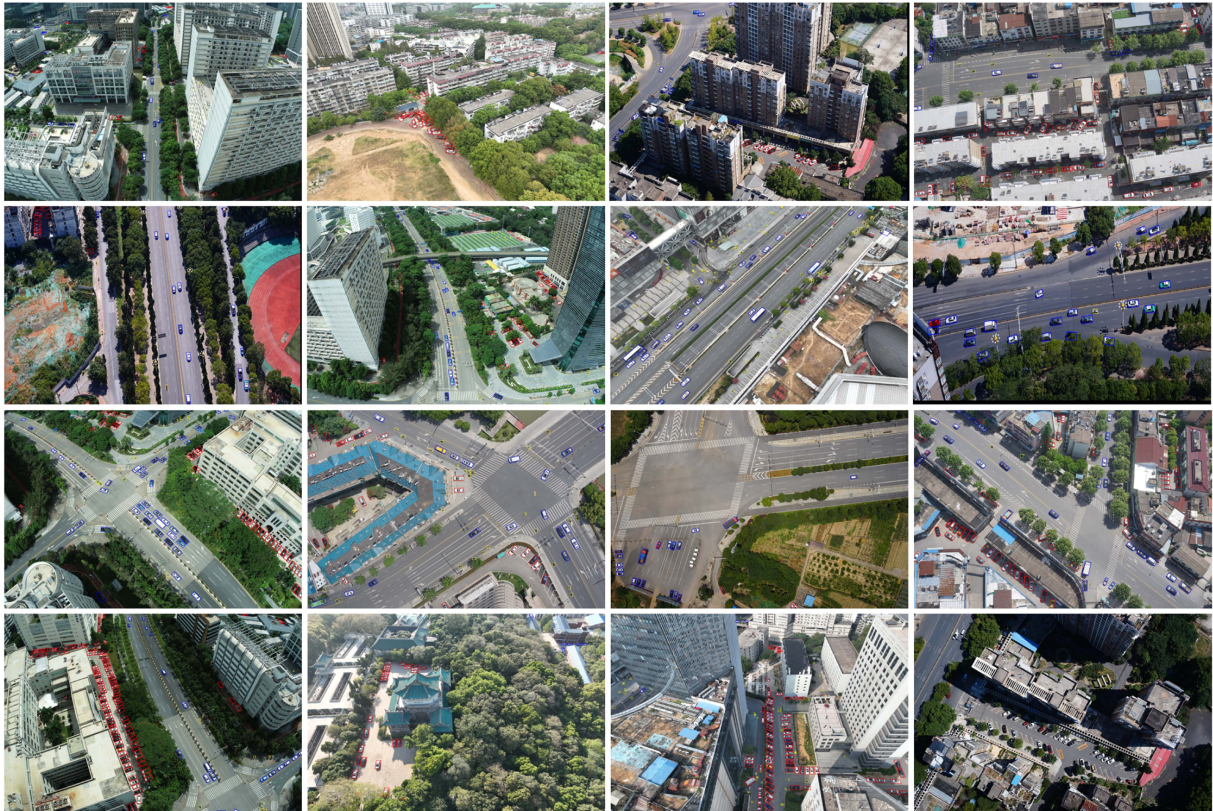


Fig. 2. Examples of different urban scenarios in PARA. The first row is the residential areas; the second row is the main roads; the third row is the crossroads; and the fourth row shows the parking lots.

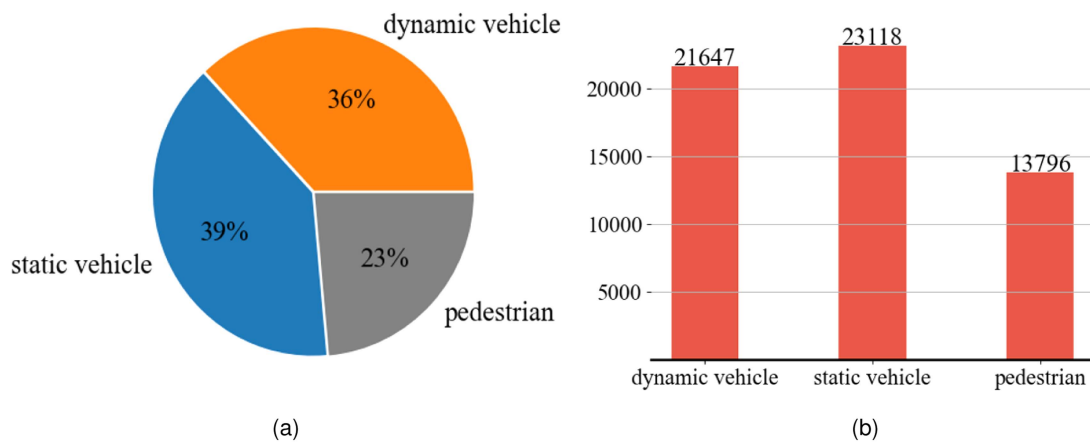


Fig. 3. Distribution of different object categories in PARA. (a) Proportion of each category in PARA. (b) Quantity of each category in PARA.

our dataset, which plays an important role in exploring the real world.

In Fig. 3(a), we show the quantity distribution of objects in PARA, with a total of 58561 annotated instances. The instances in PARA include 23118 static vehicles, 21647 dynamic vehicles, and 13796 pedestrians, and each instance has two bounding boxes, HBB and PBB. The number of each class of objects is shown in Fig. 3(b). It can be seen that static vehicles and dynamic vehicles constitute the majority of

the samples in the dataset and their distribution is relatively balanced.

C. Data Processing

Data processing is important for building a highly available dataset. Before annotation, we first manually discard a part of PARA images with poor quality, such as those blurred or broken images. Then, to ensure that the images cover vehicles of various

scales and aspect ratios, UAV images captured at different flight heights are uniformly selected. At the same time, the orientation information is also considered. The selected images in PARA are kept to contain different orientations as much as possible.

Moreover, we have developed a new annotation tool based on the labelme to outline the parallelogramlike bounding boxes for convenience. When annotating, we just need to manually find three points on the outline of a single object. Then, our tool can automatically generate a complete bounding box and compute the orientation angle of the annotated box. Therefore, we not only provide the coordinates of the original vertices, but also the orientation information of all PARA objects. The complete annotation of a single object contains the coordinates of three adjacent corners as well as the orientation degree, which ranges between 0° and 360° indicating the angle of the object head with respect to the trigonometric circle.

D. Annotation Method

In computer vision, the annotation method determines the representation of the instances and the parameters that the detectors need to learn. The compact annotated bounding boxes can contribute to separating the densely crowded objects, providing accurate semantic information to detectors. The HBB is widely used in different vision tasks, and can be denoted by (x_c, y_c, w, h) , where (x_c, y_c) and (w, h) are the center location and the size of a bounding box, respectively. Although objects in natural images can be well represented by HBB, the majority of instances with arbitrary orientations in aerial images cannot be compactly outlined by this method. In order to solve this problem, the RBB was proposed [48] for precisely locating rotated objects in aerial images. RBB additionally adds a parameter to denote the orientation of the bounding box, which can effectively separate the packed objects and reduce the background noise of the annotation in the orthographic aerial images.

While RBB can address the problem of detecting crowded objects in orthographic aerial images, this method struggles to enclose rotated objects with large geometric distortions in oblique UAV images. Objects in oblique UAV images often suffer from large perspective deformation compared with objects in natural images and orthographic aerial images. The rectangular vehicles are usually unable to remain axis-aligned under the oblique views. As a result, HBB and RBB are prone to failing to represent the accurate shape of vehicles due to the limitation of the right angle. Considering the complex backgrounds in urban scenarios and geometry distortion of vehicles, we develop a more flexible annotation method called PBB to accurately represent the vehicles in oblique UAV images. PBB is a simple and effective object representation, denoted by $\{(x_i, y_i) | i = 1, 2, 3\}$, where (x_i, y_i) represents the vertex of PBB. The fourth vertex of the PBB satisfies the following constraints: $x_4 = x_3 + (x_2 - x_1)$; $y_4 = y_3 + (y_2 - y_1)$. When annotating, we keep the long side and short side of PBBs align with the length and width of vehicles, respectively. With no restriction on the right angle, PBB can effectively encode the shape of objects with distortions caused by linear perspective in UAV imagery. To better illustrate the effectiveness of PBB,

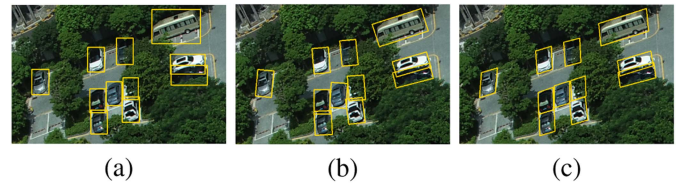


Fig. 4. Comparison of HBB, RBB, and PBB. (a)–(c) Differently annotated vehicles in crowded scenes. PBB can more compactly enclose the vehicles than HBB and RBB, which effectively reduces the noise from the background and the overlap between bounding boxes.

we show the different annotation methods applied for crowded vehicles in oblique UAV images in Fig. 4. Both of HBB and RBB introduce heavier background noise into the bounding box compared with PBB. In contrast, PBB can effectively reduce the background noise and the overlap of between crowded bounding boxes. In addition, to meet the needs of different research, we also provide manually annotated HBB of objects in PARA.

IV. PROPERTIES OF PARA

This section illustrates the main characteristics of PARA, including large scale high-resolution images, various views, differently scaled instances, and so on. Fig. 5 displays the statistical information of PARA in detail.

A. Large Scale

PARA consists of 1025 UAV images and 117 122 manually annotated bounding boxes, covering several common vehicle categories. The majority of original images in PARA have sizes of 3956×5280 pixels, 4000×6000 pixels and 3648×5472 pixels, whereas in the natural datasets the sizes of images rarely exceed 1000×1000 pixels (e.g., COCO [32] and Microsoft VOC [33]). In Fig. 5(a), we display the different resolutions of images in PARA. The high-resolution PARA images ensure a real representation of natural scenarios.

B. Various Orientations of Instances

The orientation is an important attribute of instances in object detection from UAV images. The orientation of instances not only represents the relative relationship between objects in the real world, but also has a significant impact on the feature extraction with rotation invariance. PARA provides abundant orientation information of vehicles for detectors. As shown in Fig. 5(b), the orientation angles of PARA vehicles fully distribute between 0° and 360° .

C. Multiscale Instances

The different UAV altitudes result in the different size of an instance. We adjusted the flight height of UAVs between 50 and 350 m during the collection process due to the actual sizes of vehicles in the real world do not differ significantly. The varying flight heights ensure that PARA can capture different sizes of vehicles in natural scenes, which is helpful to train a robust detector. Fig. 5(c) illustrates the notable size differences of objects in PARA.

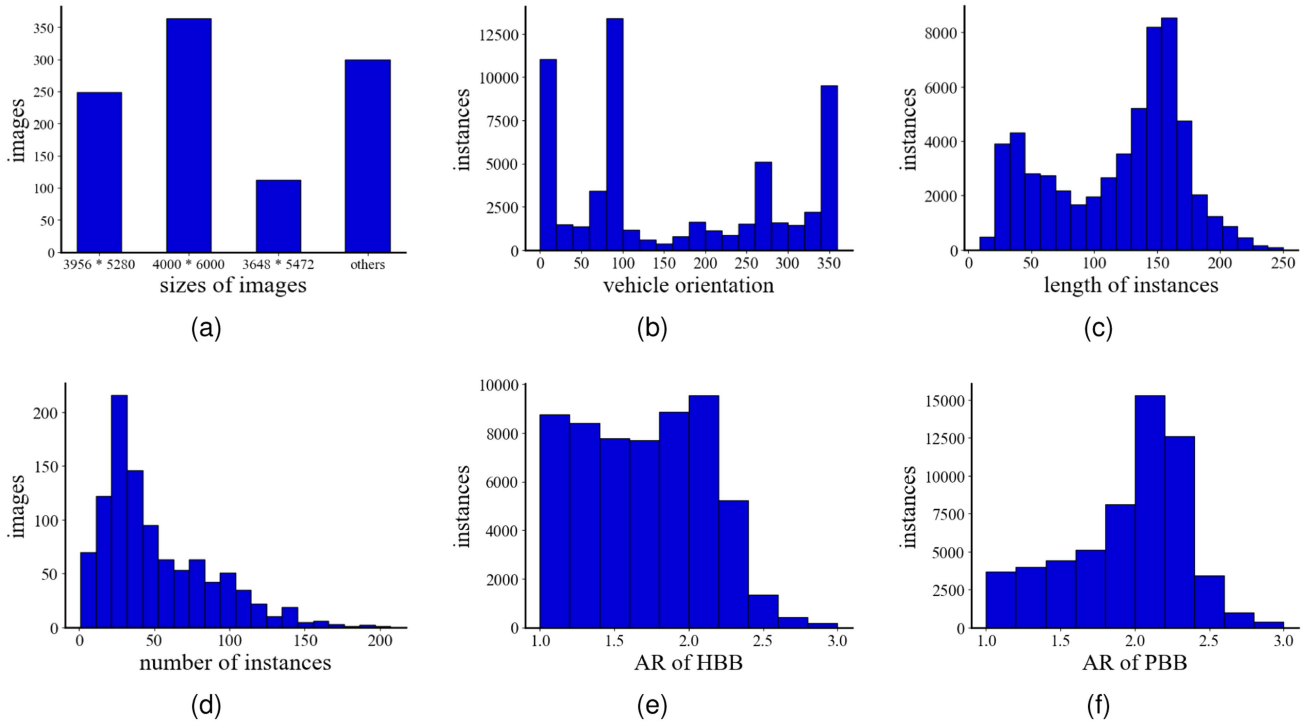


Fig. 5. Image information and object statistics for instances in PARA. AR denotes the aspect ratio. Other images refer to those in the dataset with sizes ranging from 1000 to 3000 pixels. (a) Statistics of image resolution. (b) Distribution of vehicle orientation. (c) Distribution of vehicle length. (d) Density distribution of vehicles in each image. (e) Distribution of AR for HBBs. (f) Distribution of AR for RBBs.

D. Various Density Distribution of Instances

PARA is designed specifically for vehicle detection in urban areas. We include several typical natural scenes in modern cities, such as highways, parking lots, intersections and residential areas. Different urban scenes have different background information and the presented vehicles exhibit varying density distributions in different scenarios. As shown in Fig. 5(d), a single image in PARA may contain only a few vehicles or exceed 200 number of vehicles, making PARA highly challenging.

E. Various Aspect Ratios of Instances

The aspect ratio (AR) is an important attribute of the dataset, which provides essential information about the shape and size of instances. In anchor-based detection algorithms, AR serves as an auxiliary factor that affects model design and algorithm effectiveness. For example, YOLOv3 [42] employs the k-means algorithm [70] to cluster the initial anchor sizes and ratios. We calculate two kinds of AR for all objects in PARA to provide a reference for subsequent research. Fig. 5(e) and (f) illustrates the aspect ratio of manually annotated RBBs and PBBs in PARA.

V. EXPERIMENTS

In this section, we evaluate the mainstream detectors on PARA by detecting objects with HBB and PBB, respectively. In the following, we will introduce the experimental setup, baselines of different detection tasks, experimental results, and analysis in detail.

A. Experimental Setup

In our experiment, we randomly split the sample in PARA into three parts: a training set of 511 images, a validation set of 168 images, and a testing set of 346 images, respectively. As the original PARA images are too large to be fed into the existing detectors for training, we crop them into patches of 1024×1024 pixels, with 50% overlapping between neighboring patches. Finally, we get 32 903, 11 433, and 21 809 patches for training, validation, and testing, respectively. To make a fair comparison between the baseline detectors, all models are implemented with the open-source MMDetection [71] and trained with a single GeForce RTX 3090Ti GPU. We select and evaluate Faster RCNN [36], DAB-DETR [37], Cascade RCNN [38], RTMDet [72], YOLOv3 [42], SSD [43], EfficientNet [73], RetinaNet [46], Deformable DETR [74], and FCOS [75] with ResNet50, Efficientnet B3, VGG16, CSPNeXt, DarkNet53 backbones as the baseline detectors. Specifically, we modify the original Faster RCNN [36] and RetinaNet [46] to detect vehicles with PBBs denoted by $\{(x_i, x_i), i = 1, 2, 3\}$. All the model settings are kept the same as the default setups in MMDetection [71].

B. Evaluation Metric

For the evaluation metric, we adopt mAP, the mainstream protocol in the field of object detection, to evaluate the performance of the selected baseline detectors. The mAP stands for the mean average precision (AP), which is calculated by the area under the Precision-Recall (PR) curve. PR curve can be depicted with

TABLE II
BENCHMARK OF THE STATE-OF-THE-ART ON THE HBB UNDER MAP50 AND MAP75 METRICS

Model	Backbone	AP[%](IoU=0.5)				AP[%](IoU=0.75)			
		Mean	DV	SV	P	Mean	DV	SV	P
Cascade RCNN [38]	ResNet50	79.79	88.39	86.02	64.97	63.96	76.15	74.58	41.14
FCOS [75]	ResNet50	79.06	85.92	84.20	67.07	68.66	84.73	78.99	42.25
SSD [43]	VGG16	79.55	87.32	85.32	66.00	65.11	82.35	74.04	38.94
RetinaNet [46]	ResNet50	77.32	89.39	86.05	56.53	62.61	86.37	72.82	28.65
Deformable DETR [74]	ResNet50	79.77	88.79	86.31	64.20	67.12	87.12	80.03	34.21
RTMDet [72]	CSPNeXt	79.41	88.35	85.61	64.27	65.62	83.38	74.67	38.80
YOLOv3 [42]	DarkNet53	79.44	87.68	84.97	65.67	64.57	76.09	74.02	43.62
DAB-DETR [76]	ResNet50	79.08	87.86	85.14	64.25	57.27	52.36	79.31	40.12
EfficientNet [73]	EfficientNetB3	50.27	63.04	55.15	32.63	38.86	56.38	43.83	16.37
Faster RCNN [36]	ResNet50	78.25	87.59	85.83	61.34	64.00	81.56	73.51	36.92

mAP stands for the mean average precision, higher is better.

different scores of detection precision and detection recall. The calculation of AP can be depicted as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (1)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

$$\text{AP} = \int_0^1 P(R)d(R) \quad (3)$$

where TP (true positive) refers to the number of correctly predicted bounding boxes, i.e., with an IoU score higher than the IoU threshold. FP (false positive) and FN (false negative) are the number of bounding boxes that are predicted incorrectly and not detected, respectively. The IoU threshold is often set to 0.5 and 0.75, and the corresponding mAP is denoted as mAP50 and mAP75. In addition to the commonly used mAP50 proposed in PASCAL VOC [33], we also choose mAP75 as the evaluation metric to judge whether the bounding boxes compactly enclose the targets.

C. HBB Baseline

Most of datasets for rotated object detection in aerial imagery (such as DOTA [47] and EAGLE [67]) generate HBB ground truths of instances by calculating the HBBs of RBBs. We find that the HBBs of PBBs are larger than the actual instances in oblique UAV images due to geometric distortions. Therefore, we obtain the accurate HBBs of vehicles in PARA by manual annotation.

We train the baseline models with their default hyperparameters and strategies to ensure a fair comparison. Table II displays the results of HBB prediction. Cascade RCNN [38] outperforms all the other detectors with an mAP of 79.79% for its effective training strategy based on a multistage network. The other detectors show impressive performance in our dataset with mAP50 over 77%, except for the EfficientNet [73]. We suspect that this can be attributed to the the default backbone Efficientnet-B3 in MMDetection [71], which is sensitive to the size of the input images. It is worth mentioning that the other one-stage detectors,

such as YOLOv3 [42] and SSD [43], spend more time on training in comparison with the two-stage algorithms. This may be the reason why they can achieve comparable performance with the two-stage detectors. Compared with the metric of mAP50, the mAP scores of all the detectors decrease by over 10% under the stricter metric of mAP75. Benefiting from the anchor-free strategy, FCOS [75] achieves the best performance with a mAP score of 68.66%. In the table, EfficientNet [73] presents poor performance compared with the other detectors. We find that the decrease in mAP under the metric of mAP75 is mainly caused by the difficulty in detecting pedestrians, whose sizes are too small to be accurately located. Moreover, all the detectors achieve better performance in detecting dynamic vehicles with clear backgrounds (like roads) than static vehicles with complex backgrounds.

D. PBB Baseline

Most of mainstream detectors are designed for the objects represented by HBB or RBB. It is not feasible to directly apply them as a benchmark for the PBB-based detection task. Thus, we choose and modify Faster RCNN [36] and RetinaNet [46] as the baseline for PBB detection due to their efficiency.

We modified the region proposal network (RPN) and the head of the convolution neural network (CNN) in the original Faster RCNN. Region proposals generated by the modified RPN are utilized to match the HBB of the PBB ground truths, which can be denoted by $(r_{xc}, r_{yc}, r_w, r_h)$; Then, each region proposal is fed into CNN to attach the single PBB ground truth represented by $(g_{xc}, g_{yc}, g_w, g_h, g_t, g_l)$. In detail, $(g_{xc}, g_{yc}, g_w, g_h)$ denotes the external bounding box of the PBB and (g_t, g_l) represents the relative offsets of the top vertex and the left vertex in the PBB to the left-top vertex of the external bounding box. Finally, the 6-D target vector produced by the head of the modified CNN can be written as $P = \{p_{xc}, p_{yc}, p_w, p_h, p_t, p_l\}$, where

$$p_{xc} = (g_{xc} - g_{xc})/r_w, p_{yc} = (g_{yc} - r_{yc})/r_h \quad (4)$$

$$p_w = \log(g_w/r_w), p_h = \log(g_h/r_h) \quad (5)$$

$$p_t = g_t/r_w, p_l = g_l/r_h. \quad (6)$$

TABLE III
BENCHMARK OF THE STATE-OF-THE-ART ON THE PBB UNDER MAP50 AND MAP75 METRICS

Model	Backbone	AP[%](IoU=0.5)				AP[%](IoU=0.75)			
		Mean	DV	SV	P	Mean	DV	SV	P
Cascade RCNN [38]	ResNet50	79.10	86.84	85.49	64.97	37.04	35.47	34.53	41.12
FCOS [75]	ResNet50	78.64	84.98	83.87	67.07	38.52	35.68	35.68	44.19
SSD [43]	VGG16	78.67	85.52	84.47	66.03	37.21	36.41	36.37	38.87
RetinaNet [46]	ResNet50	76.58	88.50	84.71	56.53	34.01	38.66	34.77	28.59
Deformable DETR [74]	ResNet50	78.97	87.25	85.87	63.98	36.61	37.56	36.85	35.42
RTMDet [72]	CSPNeXt	79.04	87.12	85.33	64.68	37.10	36.90	35.63	38.77
YOLOv3 [42]	DarkNet53	78.64	86.12	84.13	65.67	37.84	36.36	33.63	43.53
DAB-DETR [76]	ResNet50	78.62	86.63	84.77	64.45	36.88	35.48	34.81	40.35
EfficientNet [73]	EfficientNetB3	49.23	61.01	54.06	32.63	21.66	25.02	23.68	16.28
Faster RCNN [36]	ResNet50	77.46	86.12	84.92	61.34	35.79	34.80	35.68	36.87
Modified Faster RCNN	ResNet50	80.10	84.95	83.40	71.93	54.80	68.41	57.51	38.48
Modified RetinaNet	ResNet50	81.61	89.38	85.53	69.92	63.10	78.82	63.92	46.56

The values in bold are the best.

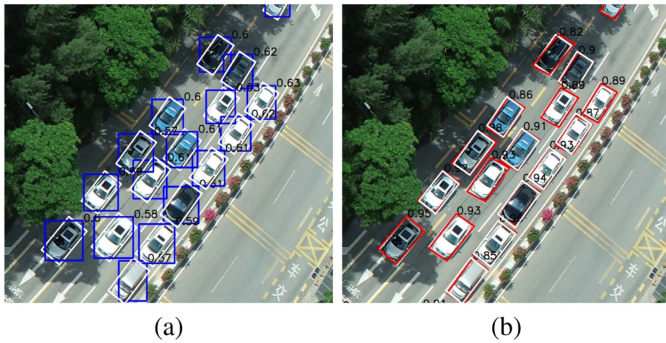


Fig. 6. Comparison of the IoU scores between HBB and PBB predicted boxes with PBB ground truths. The results show that the IoU of predicted HBB boxes is relatively low, while the predicted PBB achieves much better performance.

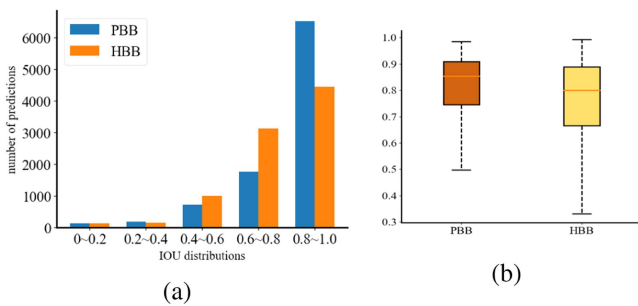


Fig. 7. Quantitative comparison between the HBB and PBB predicted boxes. (a) IoU distributions of the HBB and PBB predictions evaluated by PBB ground truths. (b) Boxplot of the IoU distribution of PBB and HBB boxes.

Similarly, we modify the original RetinaNet to regress the offsets of parallelograms to their corresponding external bounding boxes. To make a comprehensive evaluation of our modified PBB-based baselines, we also train the other mainstream detectors based on HBB annotations and evaluate the predicted results with the PBB ground truths for convenience in the PBB task.

Table III displays the results of PBB-based vehicle detection. The two PBB baselines outperform the other original state-of-the-art detectors trained with HBB. The improvement is particularly notable for the mAP75 metric, showing an increase of approximately 16% in mAP score. The results of the PBB-based detection show significant differences between PBB and HBB representation for vehicles with large geometric distortion. We also observe that the improvement in mAP score is mainly attributed to the categories of static vehicle and dynamic vehicle, with increases of around 30% and 21% in mAP, respectively. In comparison, the detection accuracy of pedestrians does not demonstrate a substantial increase due to their small size in UAV images. Overall, our findings indicate that for objects with large deformation in oblique UAV images, the PBB presentation is superior to HBB in precise and compact detection of vehicles.

VI. DISCUSSION AND ANALYSIS

In this section, we will present some interesting discussions and analysis of our experimental results. When comparing the HBB detection results in Table II with the PBB detection results in Table III, we observe that the detectors trained with HBB ground truths achieve similar performance in both HBB and PBB detection tasks under the metric of mAP50. To explain the reason behind this phenomenon, we visualize the predicted results of different tasks in Fig. 6(a) and (b). The IoU scores beside the predicted boxes in Fig. 6 can be utilized to evaluate the accuracy of predicted results, with higher values indicating better overlap between the predicted and ground truth bounding boxes. We find that the predicted PBBs in Fig. 6(a) compactly fitting ground truths, while the predicted HBBs enclose massive backgrounds. However, the predicted HBBs are also regarded as TP samples under the metric of mAP50 with a poor IoU score of 0.5 or 0.6. The results indicate that the mAP50 metric has some limitations and is unable to accurately reflect the matching degree of the predicted and ground truth boxes in oblique UAV images. We should adopt a more stringent evaluation metric like mAP75 for object detection in UAV images, where objects often suffer



Fig. 8. Visualization results of testing on PARA using well-trained original and modified Faster RCNN. (a) and (b), respectively, illustrate the predicted HBB and PBB boxes in different urban scenarios.

from large distortion. Moreover, this finding also reveals that the same detectors can achieve a completely different accuracy under the mAP75 metric, with many low-quality HBB boxes being filtered out significantly. In contrast, PBB predicted boxes achieve good performance in matching with ground truths, with IoU scores converging in the range of 0.85–0.95. It can not only fit the targets closely, but also significantly reduce overlapping with each other.

To quantitatively evaluate the matching degree of HBB and PBB predicted boxes with PBB ground truths, we collect over 9000 predicted boxes from different tasks on the validation set and visualize their distribution of IoU scores with ground truths in Fig. 7. We consider a predicted box with an IoU score above 0.8 as a positive sample. In Fig. 7(a), we observe that nearly half of HBB-predicted boxes have a high IoU score over 0.8. The IoU scores of the remaining boxes are distributed in the range of 0.4–0.8. In comparison, the majority of PBB predicted boxes have IoU scores that converge in the range of 0.8–1.0. The proportion of low-level predicted boxes is smaller in comparison to the HBB predicted boxes. In Fig. 7(b), we utilize the method of box plot to depict the distribution of IoU scores between PBB/HBB predicted boxes and the ground truths. The IoU score distribution of PBB is clustered, with a minimum around the threshold of 0.5. In comparison, the IoU distribution of HBB is scattered, showing a poor performance in matching with ground truths. Therefore, PBB is more capable of accurately representing rotated vehicles in oblique UAV images.

In Fig. 8, we select different urban scenes and compare the results between the detection results with PBB and HBB. We observe that HBB detectors classify several flower beds and trees as vehicles. This is because too much background noise contained in the bounding box disturbs the network learning. For densely arranged vehicles, the localization of objects with PBB is obviously more accurate than with HBB. HBB detectors tend

to suppress some packed detected boxes by some postprocessing operations like NMS. PBB can well address this problem by compactly enclosing crowded vehicles, resulting in better performance in crowded vehicle detection. Moreover, the loose representation of HBB causes predicted boxes to overlap with each other, while PBB can correctly reflect the actual orientation and size of vehicles in oblique UAV images. In terms of different categories, we find that the detection accuracy of static vehicles is slightly lower than that of dynamic ones. This is because static vehicles are often occluded by surrounding objects. It is relatively easy for detectors to detect dynamic vehicles with clear backgrounds.

VII. CONCLUSION

We build a large-scale dataset for vehicle detection in UAV images, namely PARA, which features a novel representation of vehicle object under oblique UAV views. Compared with the traditional annotation methods, the proposed PBB can compactly enclose the targets and provide accurate semantic information to detectors. In addition to the PBB representation, we collect a large number of high-resolution images captured in complex urban environments and manually annotate many rotated vehicles with different bounding boxes. We also evaluate the performance of several mainstream object detectors on PARA to establish a benchmark for precise vehicle detection in urban scenarios. Experimental results demonstrate that it remains challenging for detectors to accurately detect vehicles with significant deformations in complex urban scenarios.

High resolution oblique UAV images are now easily accessible with cheap drones, which provide rich information for many modern urban practical applications such as traffic monitoring, vehicle management, and urban planning. However, objects in

oblique UAV images often suffer from large perspective deformation, bringing a huge challenge for detection and analysis. We believe that the findings with the PARA dataset for compact vehicle detection can not only bring benefits addressing urban issues but also attract more attention to object detection in oblique UAV images.

REFERENCES

- [1] S. Wang, F. Jiang, B. Zhang, R. Ma, and Q. Hao, "Development of UAV-based target tracking and recognition systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3409–3422, Aug. 2020.
- [2] L. Gueguen et al., "Mapping human settlements and population at country scale from VHR images," *IEEE J. Select. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 2, pp. 524–538, Feb. 2017.
- [3] N. H. Motlagh, M. Bagaa, and T. Taleb, "UAV-based IoT platform: A crowd surveillance use case," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 128–134, Feb. 2017.
- [4] F. Bovolo, C. Marin, and L. Bruzzone, "A hierarchical approach to change detection in very high resolution SAR images for surveillance applications," *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2042–2054, Apr. 2013.
- [5] R. Min, Y. Chen, H. Wang, and Y. Chen, "DAS vehicle signal extraction using machine learning in urban traffic monitoring," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5908510.
- [6] M. Elloumi, R. Dhaou, B. Escrig, H. Idoudi, and L. A. Saidane, "Monitoring road traffic with a UAV-based system," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2018, pp. 1–6.
- [7] J. Scherer et al., "An autonomous multi-UAV system for search and rescue," in *Proc. 1st Workshop Micro Aerial Veh. Netw., Syst., Appl. Civilian Use*, 2015, pp. 33–38.
- [8] Y. Zhou, S. Chen, J. Zhao, R. Yao, Y. Xue, and A. El Saddik, "CLT-Det: Correlation learning based on transformer for detecting dense objects in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4708915.
- [9] M. R. James and S. Robson, "Mitigating systematic error in topographic models derived from UAV and ground-based image networks," *Earth Surf. Processes Landforms*, vol. 39, no. 10, pp. 1413–1420, 2014.
- [10] F. Shi, T. Zhang, and T. Zhang, "Orientation-aware vehicle detection in aerial images via an anchor-free object detection approach," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 6, pp. 5221–5233, Jun. 2021.
- [11] R. Zhang, Z. Shao, X. Huang, J. Wang, and D. Li, "Object detection in UAV images via global density fused convolutional network," *Remote Sens.*, vol. 12, no. 19, Art. no. 3140, 2020.
- [12] H. Zhang, M. Sun, Q. Li, L. Liu, M. Liu, and Y. Ji, "An empirical study of multi-scale object detection in high resolution UAV images," *Neurocomputing*, vol. 421, pp. 173–182, 2021.
- [13] Q. Li, L. Mou, Q. Xu, Y. Zhang, and X. X. Zhu, "R³-Net: A deep network for multi-oriented vehicle detection in aerial images and videos," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5028–5042, 2019.
- [14] H. Zhou et al., "Robust vehicle detection in aerial images using bag-of-words and orientation aware scanning," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 12, pp. 7074–7085, Dec. 2018.
- [15] W. Yu, T. Yang, and C. Chen, "Towards resolving the challenge of long-tail distribution in UAV images for object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2021, pp. 3258–3267.
- [16] P. Gao, T. Tian, T. Zhao, L. Li, N. Zhang, and J. Tian, "Double FCOS: A two-stage model utilizing FCOS for vehicle detection in various remote sensing scenes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4730–4743, 2022.
- [17] C. Yao, X. Bai, W. Liu, Y. Ma, and Z. Tu, "Detecting texts of arbitrary orientations in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 1083–1090.
- [18] J. Ma et al., "Arbitrary-oriented scene text detection via rotation proposals," *IEEE Trans. Multimedia*, vol. 20, pp. 3111–3122, 2018.
- [19] W. Zhang, C. Liu, F. Chang, and Y. Song, "Multi-scale and occlusion aware network for vehicle detection and segmentation on UAV aerial images," *Remote Sens.*, vol. 12, no. 11, 2020, Art. no. 1760.
- [20] G. Heitz and D. Koller, "Learning spatial context: Using stuff to find things," in *Proc. 10th Eur. Conf. Comput. Vis.: Comput. Vis.*, 2008, pp. 30–43.
- [21] F. Tanner et al., "Overhead imagery research data set—An annotated data library & tools to aid in the development of computer vision algorithms," in *Proc. IEEE Appl. Imagery Pattern Recognit. Workshop*, 2009, 2009, pp. 1–8.
- [22] K. Liu and G. Mattyus, "Fast multiclass vehicle detection on aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 9, pp. 1938–1942, Sep. 2015.
- [23] R. Krajewski, J. Bock, L. Kloeker, and L. Eckstein, "The high dataset: A drone dataset of naturalistic vehicle trajectories on German highways for validation of highly automated driving systems," in *Proc. 21st Int. Conf. Intell. Transp. Syst.*, 2018, pp. 2118–2125.
- [24] T. N. Mundhenk, G. Konjevod, W. A. Sakla, and K. Boakye, "A large contextual dataset for classification, detection and counting of cars with deep learning," in *Proc. 14th Eur. Conf.: Comput. Vis.*, 2016, pp. 785–800.
- [25] H. Zhu, X. Chen, W. Dai, K. Fu, Q. Ye, and J. Jiao, "Orientation robust object detection in aerial images using deep convolutional neural network," in *Proc. IEEE Int. Conf. Image Process.*, 2015, pp. 3735–3739.
- [26] M.-R. Hsieh, Y.-L. Lin, and W. H. Hsu, "Drone-based object counting by spatially regularized regional proposal network," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4145–4153.
- [27] Y. Zhang, Y. Yuan, Y. Feng, and X. Lu, "Hierarchical and robust convolutional neural network for very high-resolution remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5535–5548, Aug. 2019.
- [28] N. Ammour, H. Alhichri, Y. Bazi, B. Benjdira, N. Alajlan, and M. Zuair, "Deep learning approach for car detection in UAV imagery," *Remote Sens.*, vol. 9, no. 4, 2017, Art. no. 312.
- [29] Y. Xu et al., "Car detection from low-altitude UAV imagery with the faster r-CNN," *J. Adv. Transp.*, vol. 2017, 2017, Art. no. 2823617.
- [30] L. Tan, X. Lv, X. Lian, and G. Wang, "Yolov4_drone: UAV image target detection based on an improved Yolov4 algorithm," *Comput. Elect. Eng.*, vol. 93, 2021, Art. no. 107261.
- [31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [32] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. 13th Eur. Conf.: Comput. Vis.*, 2014, pp. 740–755.
- [33] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.
- [34] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.
- [35] R. Girshick, "Fast r-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1440–1448.
- [36] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-CNN: Towards real-time object detection with region proposal networks," *Adv. Neural Inf. Process. Syst.*, vol. 28, 2015.
- [37] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
- [38] Z. Cai and N. Vasconcelos, "Cascade r-CNN: Delving into high quality object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6154–6162.
- [39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.
- [40] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7263–7271.
- [41] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.
- [42] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," 2018, *arXiv:1804.02767*.
- [43] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.
- [44] C.-Y. Fu, W. Liu, A. Ranga, A. Tyagi, and A. C. Berg, "DSSD: Deconvolutional single shot detector," 2017, *arXiv:1701.06659*.
- [45] Z. Li and F. Zhou, "FSSD: Feature fusion single shot multibox detector," 2017, *arXiv:1712.00960*.
- [46] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.
- [47] G.-S. Xia et al., "DOTA: A large-scale dataset for object detection in aerial images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3974–3983.

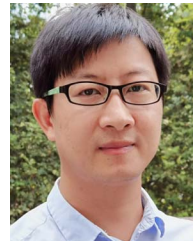
- [48] Y. Jiang et al., "R2CNN: Rotational region CNN for orientation robust scene text detection," 2017, *arXiv:1706.09579*.
- [49] L. Liu, Z. Pan, and B. Lei, "Learning a rotation invariant detector with rotatable bounding box," 2017, *arXiv:1711.09405*.
- [50] X. Yang, J. Yan, Z. Feng, and T. He, "R3det: Refined single-stage detector with feature refinement for rotating object," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 4, pp. 3163–3171.
- [51] J. Ding, N. Xue, Y. Long, G.-S. Xia, and Q. Lu, "Learning ROI transformer for oriented object detection in aerial images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2849–2858.
- [52] J. Han, J. Ding, J. Li, and G.-S. Xia, "Align deep features for oriented object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5602511.
- [53] X. Xie, G. Cheng, J. Wang, X. Yao, and J. Han, "Oriented r-CNN for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 3520–3529.
- [54] H. Schilling, D. Bulatov, R. Niessner, W. Middelmann, and U. Soergel, "Detection of vehicles in multisensor data via multibranch convolutional neural networks," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4299–4316, Nov. 2018.
- [55] J. Zhang, H. Liu, Y. Zhang, M. Li, and Z. Zhan, "Faster-transformer-ds: Multi-scale vehicle detection of remote-sensing images based on transformer and distance-scale loss," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1961–1975, 2024.
- [56] A. Robicquet, A. Sadeghian, A. Alahi, and S. Savarese, "Learning social etiquette: Human trajectory understanding in crowded scenes," in *Proc. 14th Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 549–565.
- [57] D. Du et al., "Visdrone-det2019: The vision meets drone object detection in image challenge results," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops*, 2019.
- [58] I. Bozcan and E. Kayacan, "AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 8504–8510.
- [59] D. Du et al., "The unmanned aerial vehicle benchmark: Object detection and tracking," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 370–386.
- [60] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 165, pp. 108–119, 2020.
- [61] K. Akshatha et al., "Manipal-UAV person detection dataset: A step towards benchmarking dataset and algorithms for small object detection," *ISPRS J. Photogrammetry Remote Sens.*, vol. 195, pp. 77–89, 2023.
- [62] M. Mueller, N. Smith, and B. Ghanem, "A benchmark and simulator for UAV tracking," in *Proc. 14th Eur. Conf.: Comput. Vis.*, 2016, pp. 445–461.
- [63] J. Wang, W. Guo, T. Pan, H. Yu, L. Duan, and W. Yang, "Bottle detection in the wild using low-altitude unmanned aerial vehicles," in *Proc. IEEE 21st Int. Conf. Inf. Fusion*, 2018, pp. 439–444.
- [64] L. Wen et al., "Ua-detrac: A new benchmark and protocol for multi-object detection and tracking," *Comput. Vis. Image Understanding*, vol. 193, 2020, Art. no. 102907.
- [65] S. Razakarivony and F. Jurie, "Vehicle detection in aerial imagery: A small target detection benchmark," *J. Vis. Commun. Image Representation*, vol. 34, pp. 187–203, 2016.
- [66] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogrammetry Remote Sens.*, vol. 159, pp. 296–307, 2020.
- [67] S. M. Azimi, R. Bahmanyar, C. Henry, and F. Kurz, "Eagle: Large-scale vehicle detection dataset in real-world scenarios using aerial imagery," in *Proc. IEEE 25th Int. Conf. Pattern Recognit.*, 2021, pp. 6920–6927.
- [68] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based RGB-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 10, pp. 6700–6713, Oct. 2022.
- [69] H.-Y. Lin, K.-C. Tu, and C.-Y. Li, "VAID: An aerial image dataset for vehicle detection and classification," *IEEE Access*, vol. 8, pp. 212209–212219, 2020.
- [70] J. MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Math. Statist. Probability*, 1967, vol. 1, no. 14, pp. 281–297.
- [71] K. Chen et al., "Mmdetection: Open MMLab detection toolbox and benchmark," 2019, *arXiv:1906.07155*.
- [72] C. Lyu et al., "Rtmdet: An empirical study of designing real-time object detectors," 2022, *arXiv:2212.07784*.
- [73] M. Tan and Q. L. Efficientnet, "Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.
- [74] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

- [75] Z. Tian, C. Shen, H. Chen, and T. He, "FCOS: Fully convolutional one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9627–9636.
- [76] S. Liu et al., "Dab-detr: Dynamic anchor boxes are better queries for detr," in *Proc. Int. Conf. Learn. Representations.*, 2022, *arXiv:2201.12329*.



Haitao Lv received the B.S. degree in geographic information science from the School of Geography, Nanjing Normal University, Nanjing, China, in 2020. He is currently working toward the M.S. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China.

His research interests include object detection and image inpainting.



Xianwei Zheng (Member, IEEE) received the M.S. in geographic information system and Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, China, in 2010 and 2015, respectively.

He is currently working as a Professor in computer vision and 3-D geographical information science (GIS) with Wuhan University. His research interests include indoor and outdoor scene parsing, 3-D computer vision and reconstruction, and geovisualization.



Xiao Xie received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2016.

She has been a Research Fellow with the Department of Cartography of Technical University of Munich, Munich, Germany from 2014 to 2016 and now serving as a Senior Engineer with the Key Lab of Environmental Computing and Sustainability, Liaoning, China, as well as an Assistant Professor in urban and environmental computation with the Institute of Applied Ecology, Chinese Academy of Sciences, Beijing, China. She is also a Postdoctoral Researcher with the School of Geodesy and Geomatics, Wuhan University. Her research interests include 3-D GIS and smart cities.



Xueye Chen received the B.S. degree in information engineering from Wuhan Technical University of Surveying and Mapping (now Wuhan University), Wuhan, China, and the M.S. degree in software engineering from Huazhong University of Science and Technology, Wuhan, in 1995 and 2007, respectively.

He primarily engages in technical research, application development of geographic information systems (GIS), remote sensing, digital government, and smart city technologies.



Hanjiang Xiong received the B.S. degree from the School of Remote Sensing and Engineering, Wuhan University of Surveying and Mapping, Wuhan, China, in 1995 and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing (LIESMARS), Wuhan University, Wuhan, in 2002.

He worked as a Visiting Scholar with Queensland University of Technology, Brisbane City, QLD, Australia for three months in 2011. He is currently

working as a Full Professor in 3-D GIS, Wuhan University. His current research interests include geospatial data management, 3-D visualization, augmented reality, and indoor and outdoor GIS.